Exploiting a comparability mapping to improves bi-lingual data categorization: a three-mode data analysis perspective

Pierre-Francois Marteau pierre-francois.marteau@univ-ubs.fr
IRISA (UMR CNRS 6074), Universite de Bretagne Sud, 56000 Vannes, France

Guiyao Ke yannick.crystal@gmail.com
IRISA (UMR CNRS 6074), Universite de Bretagne Sud, 56000 Vannes, France

Abstract

We address in this paper the co-clustering and co-classification of bilingual data laying in two linguistic similarity spaces when a comparability measure defining a mapping between these two spaces is available. A new approach that we can characterized as a three-mode data analysis scheme, is proposed to mix the comparability measure with the two similarity measures. Our aim is to improve jointly the accuracy of classification and clustering tasks performed in each of the two linguistic spaces, as well as the quality of the final alignment of comparable clusters that can be obtained. We used first some purely synthetic random data sets to assess our formal similarity-comparability mixing model. We then propose two variants of the comparability measure that has been defined by (Li & Gaussier, 2010) in the context of bilingual lexicon extraction to adapt it to clustering or categorizing tasks. These two variant measures are subsequently used to evaluate our similarity-comparability mixing model in the context of the co-classification and co-clustering of comparable textual data sets collected from Wikipedia categories for the English and French languages. Our experiments show clear improvements in clustering and classification accuracies when mixing comparability with similarity measures, with, as expected, a higher robustness obtained when the two comparability variant measures that we propose are used. We believe that this approach is particularly well suited for the construction of thematic comparable corpora of controllable quality.

1. Introduction

Parallel corpora are sets of tuples of aligned documents that are formed with texts placed alongside with their translation(s). If such resources are of great utility in particular in the field of assisted translation or multilingual information retrieval, they are expensive to develop and often difficult to transpose from a specialty domain to another. The notion of comparable corpora has emerged in the nineties to palliate this lack of versatility and expensiveness and to offer avenues to a wider scope of applications such as multilingual terminology extraction, multilingual information retrieval or knowledge engineering (Baker, 1996), (EAGLES, 1996). However, the notion of comparability between documents expressed in different languages is not easy to introduce: it is widely admitted that two documents in different languages are comparable when they share analogous criteria of composition, genre and topics. The term of comparable corpora was introduced by (Fung & Yee, 1998), (Munteanu, Fraser, & Marcu, 2004) and remains quite subjective. (Déjean & Gaussier, 2002) proposed a quantitative definition of the concept of comparability according to which "Two corpora in two languages $\mathcal{L}_1$ and $\mathcal{L}_2$ are called comparable if there is a significant

sub-part of the vocabulary of the $\mathcal{L}_1$ language corpus, respectively $\mathcal{L}_2$ language corpus, whose translation is in the corpus of language $\mathcal{L}_2$, respectively $\mathcal{L}_1$." (Li & Gaussier, 2010) have then derived a quantitative measure that is based on a bilingual translation dictionary. This measure consists primarily in counting the presence of the translations of dictionary entries that occur in the paired documents. It depends in a non-explicit way upon jointly the coverage of the bilingual translation dictionary and the studied corpora themselves.

This comparability measure defined for bilingual corpora indeed applies when dealing with monolingual documents that partition in two distinct linguistic spaces, as far as a bilingual dictionary connecting the two spaces is available. At a document level we thus face a situation where monolingual similarity measures exist in each linguistic space that are potentially linked by a comparability measure. In the scope of the construction of thematic comparable corpora, this leads to address the co-classification or co-clustering of bilingual data since we are targeting the mapping of highly *comparable* clusters of documents that are furthermore thematically coherent in each linguistic space, i.e. characterized by a high *intra-similarity*. We confront such situation when harvesting multilingual data from the web for instance. With the need for comparable resources getting pressing, approaches that exploit consistently similarity and comparability measures are becoming particularly useful.

There is apparently no existing direct method available to map comparable clusters of documents that lay in two different linguistic spaces. Nevertheless, there exist some work which is somehow related to this problem, like biclustering, co-clustering, or two-mode clustering introduced by (Mirkin, 1996) and (Van Mechelen I, 2004). However, these works are mainly relevant to the clustering of the rows and columns (instances and features axes) of a given matrix and does not fit with the sort of three-mode categorization or clustering we are facing.

Recently, (Jagarlamudi, Daumé, & Udupa, 2011a), (Jagarlamudi, Udupa, Daumé, & Bhole, 2011b) have developed quite successfully a supervised method that learns interlingual representations from aligned training documents. They exploit word association measures and bilingual dictionary to remove noisy pairs of aligned documents. In (Amini & Goutte, 2010) the authors proposed to learn a co-classification from multi-lingual corpora, based on a co-regularization of the categories in order to maintain a consistency of the categorization process across languages. (Li, Gaussier, & Aizawa, 2011) have proposed a solution for clustering bilingual corpora by using the comparability measure only.

However, if our approach also seeks the joint clustering or classification of data that lay in two distinct linguistic spaces, it aims at exploiting, in conjunction with a comparability mapping existing between the two spaces, *native* similarity measures (a native similarity has to be understood as any quantitative intra-language similarity measure, such as a cosine similarity measure) existing within these two linguistic spaces. More precisely, the proposed approach is lying between the work reported in (Amini & Goutte, 2010) and (Li et al., 2011). It exploits directly, i.e. without any learning phase, the comparability measure that maps the two linguistic spaces to provide new similarity measures that combine *native* similarity measures with a similarity measure that is *induced* by the comparability mapping.

Thus the approach that we develop in the following sections only rely on a bilingual dictionary and does not assume that any aligned data preexist as learning data. Indeed, this approach could be enriched using a feature-extraction technique, such as the one proposed

in (Vu, Aw, & Zhang, 2009) for instance, to align bilingual documents that have a similar content.

After introducing our main motivations, we develop a straightforward mixing model to combine similarity and comparability measures in an efficient way that allows for the development of consistent co-clustering and co-classification of comparable data and assess it on purely synthetic data. We then address the concept of comparability for mapping bilingual textual data, and define, from the original measure proposed by (Li et al., 2011), two alternative variant measures to overcome some limitation of the original measure. To assess the proposed approach on real textual data, we then detail an experimentation based on a subset of comparable documents collected from some Wikipedia categories. Basically, we evaluate jointly the three tested comparability measures and the proposed similarity-comparability mixing model in the scope of co-classification and co-clustering of bilingual data. Finally we discuss our results and draw some perspectives.

2. Motivations: similarity spaces connected by a comparability mapping

When confronting with complex data one may encounter situations where two distinct spaces $\mathcal{S}$ and $\mathcal{S}'$, in which preexist *native* similarity measures $S_{\mathcal{S}}$ and $S_{\mathcal{S}'}$, are interconnected by a mapping $C_{\mathcal{S}\mathcal{S}'}$. Figure 1 gives an example of such situation. This is the case when considering comparable corpora that are composed with texts written in at least two distinct languages. For such data, a bilingual dictionary allows for the construction of a comparability measure (Li & Gaussier, 2010) yielding to the definition of a comparability mapping (Marteau & Ménier, 2013) that links the two sets of comparable documents. More generally speaking, such case arises in situations where heterogeneous but analogous data is available, through different sources, in different formats, or characterized using different sets of descriptors, or comply to different semantic models such as heterogeneous ontologies for instance, etc. The principle of mapping heterogeneous but comparable data that we address is quite general since it takes the form of any bipartite weighted undirected graph. We call it a comparability mapping. Hence, a comparability mapping establishes a bi-directional connection between the elements of the two similarity spaces that could be used to challenge native similarity measures (or distances) defined in the two spaces. By doing so, we introduce a kind of three mode data analysis scheme: the two first modes are associated to the two *native* similarity spaces, while the third mode is related to the comparability mapping itself that links these two spaces.

As an example, in figure 1, two discrete sets of elements $\mathcal{S}$ and $\mathcal{S}'$ are presented. We suppose that the notion of *native* similarity between elements of these sets is defined, we call them respectively $S_{\mathcal{S}}(.,.)$: $\mathcal{S} \times \mathcal{S} \to \mathbb{R}$ and $S_{\mathcal{S}'}(.,.)$: $\mathcal{S}' \times \mathcal{S}' \to \mathbb{R}$.
Furthermore, the two sets are point-wisely connected by a mapping defined by a comparability measure $C_{\mathcal{S}\mathcal{S}'}(.,.)$: $\mathcal{S} \times \mathcal{S}' \to \mathbb{R}$. This mapping that takes the form of a bipartite graph is a comparability mapping. The edges of this graph are bidirectional and weighted by a real value that can be bounded into $[-1, 1]$.

The main idea that we develop in this article is that of similarity *induced* by a comparability mapping: in other words, if two elements in the set $\mathcal{S}$ are mapped to a same subset of elements in the set $\mathcal{S}'$, then their similarity should be important (and vice versa). *a contrario*, if two elements in the set $\mathcal{S}'$ are mapped to disjoint subsets of elements in the set

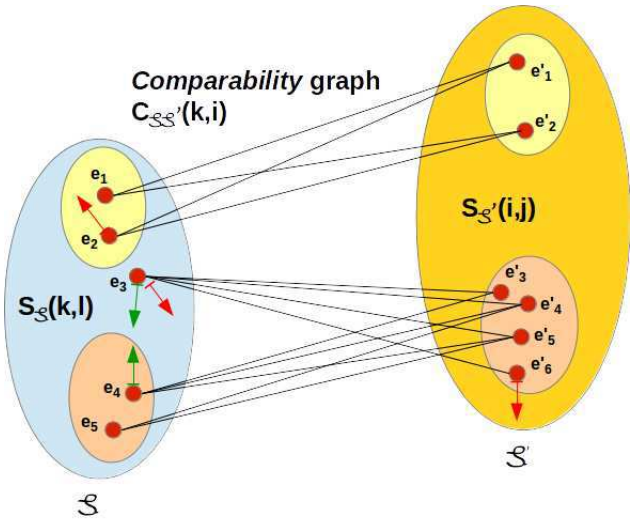

Figure 1: Two similarity spaces connected by a comparability mapping.

$\mathcal{S}$, then their similarity should be small (and vice versa). Thus, in figure 1, from the point of view of the similarity derived from the comparability mapping alone, element e_3 should move away from element e_2 to get closer to elements e_4 and e_5. Similarly, element e'_6 should move away from elements $e'3$, e'_4 and e'_5. The expected utility of such a similarity *induced* by a comparability mapping is a kind of noise filtering capability. When exploited in conjunction with *native* similarity measures in $\mathcal{S}$ and $\mathcal{S}'$ a fusion of complementary sources of knowledge is achieved that could help building more robust similarity functions into $\mathcal{S}$ and $\mathcal{S}'$ spaces. The noise in question could have many sources, in particular it could be inherent to the representational models of the element themselves due to a lack of knowledge, e.g. lack of structural variability, data heterogeneity, semantic ambiguities, etc.

3. Combining similarity and comparability: a three-mode analysis scheme

3.1 Similarity measure *induced* by a comparability mapping

In this line of work, (Marteau & Ménier, 2013) proposed an algorithm, *Hit-ComSim*, to iteratively construct the concept of similarity *induced* by a comparability bipartite graph. Unfortunately, this algorithm does not scale well due to its high algorithmic complexity (in $O(N^4)$). We propose here a much more straightforward approach that consists in exploiting directly the comparability matrix constructed from the two bilingual finite collections of documents.

Let us consider $\mathcal{S}$ and $\mathcal{S}'$ two collections of documents belonging to two distinct linguistic spaces ($\mathcal{L}$ and $\mathcal{L}'$ respectively) in which two *native* similarity measures $S_{\mathcal{S}}$ and $S_{\mathcal{S}'}$ are defined. Let $C(.,.) : S_{\mathcal{S}} \times S_{\mathcal{S}'} \rightarrow \mathcal{R}$ be the comparability function that maps the two finite collections or equivalently that defines a weighted bipartite graph between the two linguistic spaces. The two similarity functions $S_{\mathcal{S}}$, $S_{\mathcal{S}'}$ and the comparability measure C allows for the definition of the following three-mode analysis scheme.

We define the similarity measures *induced* by the comparability mapping C as the following two normalized (in $[-1,1]$) measures respectively noted $S_{\mathcal{S}_1,C}$ and $S_{\mathcal{S}',C}$:

$\forall(d_i,d_j)\in\mathcal{S}^2$ and $\forall(d'_i,d'_j)\in\mathcal{S}'^2$

$$\begin{aligned} S_{\mathcal{S},C}(d_i,d_j) &= \frac{CC^T(i,j)}{\sqrt{CC^T(i,i)CC^T(j,j)}} \\ S_{\mathcal{S}',C}(d'_i,d'_j) &= \frac{C^TC(i,j)}{\sqrt{C^TC(i,i)C^TC(j,j)}} \end{aligned} \tag{1}$$

The interpretation of the similarity measures that are *induced* by a comparability mapping C is straightforward. First, considering each row i of the C matrix as a feature vector that characterizes document $d_i\in\mathcal{S}$, for any $(d_i,d_j)\in\mathcal{S}$, $CC^T(i,j)$ can be interpreted as an inner product between the two feature vectors representing d_i and d_j respectively. Then, $S_{\mathcal{S},C}(d_i,d_j)$ is nothing but a cosine similarity between documents d_i and d_j based on the comparability mapping only.

Similarly, considering each column i of the C matrix as a feature vector that characterizes document $d'_i\in\mathcal{S}'$, $S_{\mathcal{S}',C}(d'_i,d'_j)$ is nothing but a cosine similarity between documents d'_i and $d'_j\in\mathcal{S}'$ based on the comparability mapping only.

3.2 Mixing *native* similarity and *induced* similarity

The comparability/similarity mixing model that we propose is a simple linear combination of the *native* and *induced* similarity measures defined in each linguistic space. Basically we use a single parameter $\alpha\in[0,1]$ to combine linearly the two measures as follows

$$\begin{aligned} S'_{\mathcal{S}}(d_i,d_j) &= \alpha S_{\mathcal{S},C}(d_i,d_j)+(1-\alpha)S_{\mathcal{S}}(d_i,d_j) \\ S'_{\mathcal{S}'}(d'_i,d'_j) &= \alpha S_{\mathcal{S}',C}(d'_i,d'_j)+(1-\alpha)S_{\mathcal{S}'}(d'_i,d'_j) \end{aligned} \tag{2}$$

Since the *induced* similarity measures are normalized into the interval $[-1,1]$, we advocate using a cosine similarity as *native* similarity measures in the two connected linguistic spaces such that the mixed similarity measures defined by equation 2 are consistent.

Finally, as this model mixes two sources of native similarity with the induced similarity measures that are directly derived from the comparability mapping, it implements the so-called three-mode data analysis scheme that we were referring to in the motivation section.

4. Experimenting on synthetic data

To evaluate the effectiveness of the proposed similarity-comparability mixing model, we generated 20 distinct tests by randomly defining:

- two similarity spaces $\mathcal{S}$ and $\mathcal{S}'$,
- a categorization of the elements within each of these spaces,
- the comparability mapping between them.

The algorithm 1 describes the way each of these 20 tests is generated. The variance parameters V_s and V_c are set up such that the categories are significantly overlapping making the classification problems difficult enough. To put more discriminative weight on the *native* similarity measures, the variance V_c associated to the comparability mapping matrix that is used to provide the *induced* similarity measures is three times the variance V_s used for producing the *native* similarity measures.

Algorithm 1 Random generation of two *native* similarity spaces connected by a comparability mapping. The algorithm provides two random similarity matrices $S_{\mathcal{S}}$ and $S_{\mathcal{S}'}$, the random comparability mapping matrix $C_{\mathcal{S},\mathcal{S}'}$, two sets of comparable clusters associated to a cluster map, $mapC$, also randomly defined on spaces $\mathcal{S}$ and $\mathcal{S}'$.

// $Randn(n, m)$ returns an n-by-m matrix containing pseudo-random values drawn
// from a normal distribution with mean zero and standard deviation one.
// $Randn()$ returns a single value from the previous distribution.
0) $V_s = 1.0$; $V_c = 3.0$;
1) Randomly select the number of clusters $n^c_{\mathcal{S}}$ (resp. $n^c_{\mathcal{S}'}$) in $\mathcal{S}$ (resp. $\mathcal{S}'$) from the set $\{3, ..., 18\}$;
2) For each cluster c_k in $\mathcal{S}$ (resp. c'_l in $\mathcal{S}'$) randomly select the number of elements in c_k, $|c_k|$ (resp.c'_l, $|c'_l|$) from the set $\{20, .., 40\}$;
3) For each pair of elements (e_i, e_j) in $\mathcal{S}^2$ (resp. $\mathcal{S}'^2$)
if e_i and e_j belong to the same cluster **then**
  $S_{\mathcal{S}}(e_i, e_j) = 0.5 + V_S * Randn()$;
  resp. $S_{\mathcal{S}'}(e_i, e_j) = 0.5 + V_S * Randn()$;
else
  $S_{\mathcal{S}}(e_i, e_j) = -0.5 + V_S * Rand()$;
  resp. $S_{\mathcal{S}'}(e_i, e_j) = -0.5 + V_S * Rand()$;
end if
4) $mapC = Randn(n^c_{\mathcal{S}}, n^c_{\mathcal{S}'})$;
$J = 0$
for $k = 1 : n^c_{\mathcal{S}}$ **do**
  $I = 0$;
  for $l = 1 : n^c_{\mathcal{S}'}$ **do**
    for $i = 1 : |c'_l|$ **do**
      for $j = 1 : |c_k|$ **do**
        $C_{\mathcal{S},\mathcal{S}'}(I + i, J + j) = randn() * Vc + mapC(l, k)$;
      end for
    end for
    $I = I + |c'_l|$;
  end for
  $J = J + |c_k|$;
end for

Table 1 gives for each similarity spaces $\mathcal{S}$ and $\mathcal{S}'$ the number of elements and the number of clusters for each of the 20 tests.

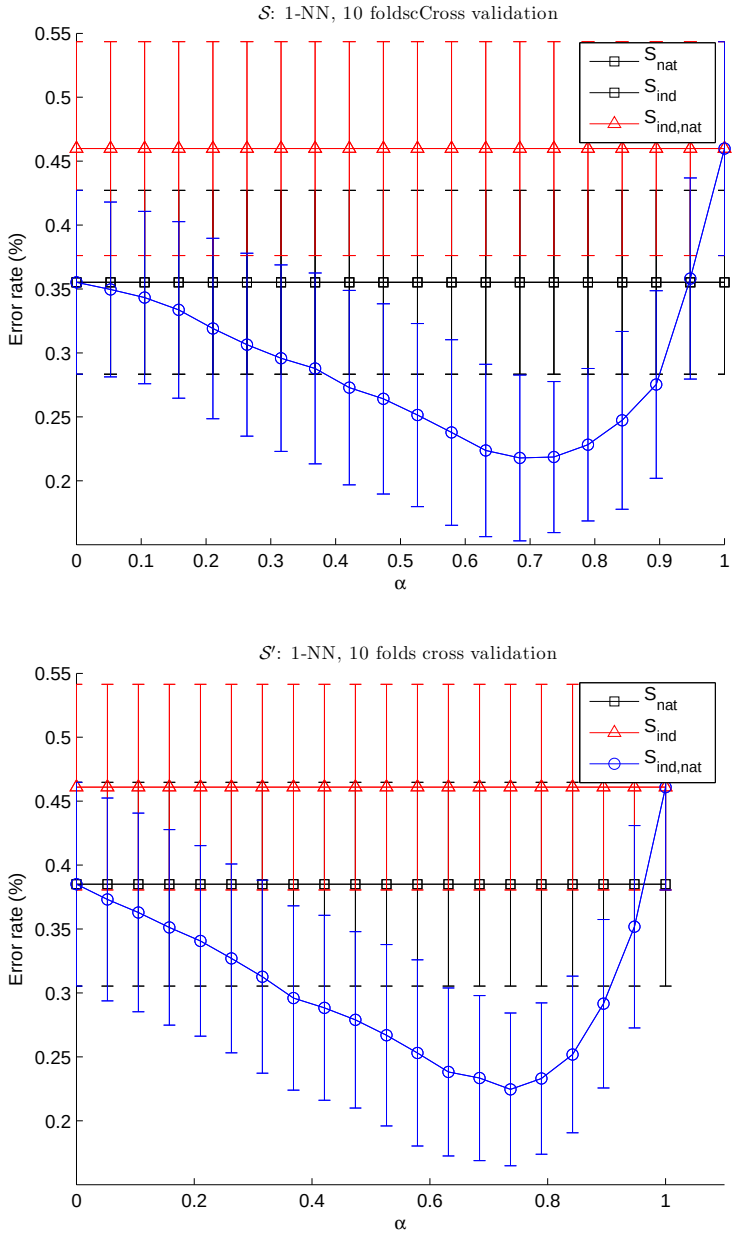

Figure 2: Comparability/similarity mixing effect on the 1-NN classification task, according to a 10-fold cross validation procedure. The error rates in % are given when the *native* similarity alone is used (black square curve) when the *induced* similarity alone is used (red triangle curve) and when the mixing model is used as a function of parameter $\alpha \in [0, 1]$ (blue circle curve). Top $\mathcal{S}$, bottom $\mathcal{S}'$ similarity spaces

$\mathcal{S}$		$\mathcal{S}'$		$\mathcal{S}$		$\mathcal{S}'$	
#clusters	#elements	#clusters	#elements	#clusters	#elements	#clusters	#elements
15	491	7	189	15	411	11	314
12	355	5	148	15	524	12	343
16	485	10	309	6	169	17	488
16	516	8	243	8	223	14	420
9	287	17	486	18	546	5	147
8	247	11	322	13	415	12	385
18	540	14	405	11	347	16	499
14	445	12	386	11	366	6	191
8	246	17	540	17	547	17	528
18	565	4	136	5	160	16	456

Table 1: Number of clusters and number of total elements in each synthetic similarity spaces $\mathcal{S}$ and $\mathcal{S}'$ for the twenty tests used for this experiment.

To evaluate the effectiveness of a 1-NN classification as the mixing parameter α varies, we use the classification error rate measure. The mean and the variance for this measure are estimated on the basis of the 20 tests and the 10-fold cross validation.

Figure 2 gives the mean and variance of the error rates for the 20 tests obtained when a 10 fold cross validation is performed using a 1-NN classifier. As shown in this figure the 1-NN classification using the *induced* similarity measures alone performs the worse, which was expected since the variance on the comparability mapping is three times the one used to generate the native similarity measures. The error rate is thus 46%in $\mathcal{S}$ and in $\mathcal{S}'$ when the *induced* similarity measures alone are used and 35% in $\mathcal{S}$ and 39% in $\mathcal{S}'$ when the *native* similarity measures alone are used.

The effect of mixing *native* and *induced* similarity measures is strongly effective on these synthetic data sets since for both spaces the error rates drop below 25% and reach a minimum when $\alpha = 0.75$. Note that the variance of the mean error decrease slightly when the mixing parameter α is around this optimal value 0.75. This experiment shows that even when the *native* and *induced* similarity measures are significantly noisy, the combination of the two sources of information allows for a significant reduction of the noise.

This is precisely this effect that we would like to show on real bilingual comparable data, when the comparability mapping is elaborated from a bilingual lexicon.

5. Variations around a quantitative comparability measure for bilingual texts

5.1 Comparability measure by Li and Gaussier (C_{LG})

The quantitative comparability measure proposed by (Li & Gaussier, 2010) is based on the simple counting of *word translation connections* that exist between two corpora in different languages according to a translation lexicon. Formally, let $\mathcal{S}_1$ and $\mathcal{S}_2$ be two corpora ex-

pressed respectively in language $\mathcal{L}_1$ and $\mathcal{L}_2$. This comparability measure is formally defined as:

$$C_{LG}(\mathcal{S}_1, \mathcal{S}_2) = \frac{\sum_{w_1 \in W\mathcal{S}_1 \cap WD_1} \sigma(w_1) + \sum_{w_2 \in W\mathcal{S}_2 \cap WD_2} \sigma(w_2)}{|W\mathcal{S}_1 \cap WD_1| + |W\mathcal{S}_2 \cap WD_2|} \quad (3)$$

where: $W\mathcal{S}_i$, $i \in \{1, 2\}$ is the lexicon in language $\mathcal{L}_i$ associated with the corpus $\mathcal{S}_i$; WD_i is the set of entries for language $\mathcal{L}_i$ into the bilingual dictionary that occur in $W\mathcal{S}_i$; $\sigma(w_i)$ is an indicator function that takes the value 1 if at least one potential translation of the term $w_i \in W\mathcal{S}_i$ in language $\mathcal{L}_i$ exists in the vocabulary associated with the corpus of the other language, 0 otherwise.

This measure was originally designed for a bilingual lexicon extraction purposes, and not for the clustering or categorization of textual data. Hence, the authors did not incorporate any term weighting since it is *a priori* irrelevant for a lexicon extraction task. However, if their definition is in line with a general definition of comparability such as the one given in introduction, the lack of term weighting is questionable when addressing a clustering/categorization task. The two variants that we propose hereinafter introduce a term weighting based on the number of term occurrences to specifically adapt the measure defined by (Li & Gaussier, 2010) to clustering or categorizing tasks.

5.2 Enrichment of the C_{LG} measure

The C_{LG} measure proposed by Li and Gaussier (eq.3) takes account of neither the number of occurrences of the lexical entries in the documents nor their number of translations into the paired documents. The binary presence or absence of joint translation entries that is modeled by the indicator function $\sigma(w_i)$ is a strong feature that may affect the average comparability between pairs of documents. This could be the case when addressing corpora for which frequency of lexical entries helps discriminating between genres and topics. We propose the following two similar variants of the C_{LG} measure that explicitly propose to go beyond the presence or absence of joint translations, conjecturing that this improvement will produce a positive effect in certain situations and tasks.

5.2.1 First variant : C_{VA_1}

The first variant symmetrically exploits (from the stand point of $\mathcal{L}_1$ and $\mathcal{L}_2$ languages) the following three elements: the number of occurrences of entries w taken into the vocabulary of the first language corpus, the number of their translations in the bilingual dictionary and the presence of at least one of their translations in the vocabulary of the second language corpus.

Let $A_{1|2}$, A_1, $A_{2|1}$, A_2 be defined as follows:

$$A_{1|2} = \sum_{w_1 \in WS_1 \cap WD_1} \left(\frac{tf(w_1, \mathcal{S}_1)}{\tau(w_1, WD_1)} \cdot \sigma(w_1) \right)$$
$$A_1 = \sum_{w_1 \in WS_1 \cap WD_1} \left(\frac{tf(w_1, \mathcal{S}_1)}{\tau(w_1, WD_1)} \right)$$
$$A_{2|1} = \sum_{w_2 \in WS_2 \cap WD_2} \left(\frac{tf(w_2, \mathcal{S}_2)}{\tau(w_2, WD_2)} \cdot \sigma(w_2) \right)$$
$$A_2 = \sum_{w_2 \in WS_2 \cap WD_2} \left(\frac{tf(w_2, \mathcal{S}_2)}{\tau(w_2, WD_2)} \right)$$

where $tf(w_i, \mathcal{S}_i)$ is the number of occurrences of entry w_i in the corpus $\mathcal{S}_i$ expressed in language $\mathcal{L}_i$, $i \in \{1, 2\}$; $\tau(w_i, WD_i)$ is the number of translations of entry w_i of the corpus $\mathcal{S}_i$ in the dictionary WD_i; $\sigma(w_i)$ is defined as above.

$$C_{VA_1} = \frac{1}{2} \cdot \left(\frac{A_{1|2}}{A_1} + \frac{A_{2|1}}{A_2} \right) \tag{4}$$

5.2.2 Second variant : C_{VA_2}

This second variant is very similar to the previous one. It distinguishes mainly on the way the measure is symmetrized. Basically the first variant relates to a geometric mean while the second variant relates to an arithmetic mean.

$$C_{VA_2} = \frac{A_{1|2} + A_{2|1}}{A_1 + A_2} \tag{5}$$

6. Experimenting on textual bilingual data

We have collected the assessment corpora from 21 Wikipedia categories, from English (EN) and French (FR) languages. It originally consists of 154828 documents in total with 87793 English documents and 67035 French documents categorized in 21 categories, taken from existing Wikipedia categories. Since such corpus is thematically very large, corresponding similarity and comparability matrices are basically very sparse. To avoid the algorithmic complexity behind the calculation of the induces similarity matrices ($O(N^3)$), we proceeded as follows which drastically reduces the sparsity of our matrices:

1. For each class and each language, we evaluate firstly the intra-language similarity matrices, using a cosine similarity based on a $tf - idf$ weighting,
2. secondly, we prune these intra-language similarity matrices using a threshold (typically 0.5) and order the documents according to their number of remaining neighbors (with whom they share a similarity above the threshold).
3. by keeping for each language the best hundred documents, we get a refined comparable bilingual corpus.

4. Finally, to complexify the experiment, we enrich this corpus by adding, for each language, and for each class, 50% of the initial number of documents. These added documents are randomly drawn from the initial 21 Wikipedia categories.

Each Wikipedia article is then represented by its plain textual content. tags and hyperlink have thus been removed. This Wikipedia corpus[1] contains 5822 documents in total, and is composed with 2745 French documents and 3077 English documents distributed into the 21 categories as listed in Table 2.

EN categories	# doc	FR categories	# doc	EN categories	# doc	FR categories	# doc
Astronomy	151	Astronomie	123	Movie	151	Film	151
Biology	151	Biologie	115	Music	151	Musique	151
Economy	144	Economie	151	Skating	151	Patinage	151
Food	147	Nourriture	4	Heritage	151	Patrimoine	151
Football	151	Football	151	Politics	151	Politique	151
Genetics	82	Génétique	151	Religion	150	Religion	133
Geograpphy	139	Geographie	151	Rugby	151	Rugby	151
Computer	151	Ordinateur	151	Health	151	Santé	63
Literature	150	Littérature	151	Sculpture	151	Sculpture	151
Mathematics	151	Mathématique	63	Tennis	151	Tennis	151
Medicine	151	Médecine	130				

Table 2: Composition of the comparable bilingual corpus extracted from Wikipedia (EN: English, FR: French)

This corpus has been lemmatized using the TreeTagger (Schmid, 1994) (Schmid, 2009). Stoplists for French and English languages have been used and the term frequencies (tf) for each vocabulary entry/document pair have been evaluated, as well as the inverse document frequencies idf (Spärck Jones, 1972) that were estimated on the corpus. Each Wikipedia article is finally represented by a tf-idf weighted vector according to the classical vector space model (Salton, Wong, & Yang, 1975).

6.1 Bilingual dictionary

To estimate the quantitative comparability between a pair of Enlish/French documents we have used the bilingual dictionary available at ELRA under reference ELRA-M0033. This dictionary contains 243,580 pairs of lexical entries in French and in English, which decompose into 110,541 lexical entries in English and 109,196 lexical entries in French.

The influence of the dictionary coverage rate has been partially studied in (Li & Gaussier, 2010) and (Ke, Marteau, & Ménier, 2014). It is shown that, for all three comparability measures C_{LG}, C_{VA_1} and C_{VA_2}, the correlation of these measures with a gold standard comparability measure reference degrades when the dictionary coverage rate relatively to the corpus lexicon decreases. We do not address this issue in this paper, keeping in mind

1. The Wikipedia corpus is available at `http://people.irisa.fr/Pierre-Francois.Marteau/Corpora/Wikipedia_21classes.zi`

that an enrichment of the bilingual dictionary by including in particular domain dependent bilingual terminology entries would likely greatly improve our results.

6.2 Evaluation measures

The performance of the 1-NN classifier is evaluated using the classification error rate estimate using a 10-fold cross validation. The performance of the tested clustering algorithms are also evaluated by comparing the predicted label for each document with its *true* label. The accuracy (AC) and normalized mutual information (NMI) measures are used to evaluate the clustering performance (Wei Xu & Gong, 2003). As an internal evaluation scheme for estimating the quality of the clustering obtained in each linguistic space, we also use the Davies–Bouldin index (DB) (Davies & Bouldin, 1979) which roughly measures the quotient of intra and inter cluster average similarity measures.

The accuracy (AC) measure is defined as follows: it measures the fraction of documents that are correctly labeled, assuming a one-to-one correspondence between true categories and assigned clusters. Let p denote any possible permutation of index set of clusters and *true* categories. The Accuracy is thus defined as

$$AC = \frac{1}{N} MAX_p \sum_{i=1\cdots K} n_{i,p(i)} \tag{6}$$

where $n_{i,p(i)}$ denotes the number of documents shared by class i and cluster $p(i)$, K is the number of categories and clusters, and N is the total number of documents.

The NMI measure between the *true* clustering $\mathcal{C}$ and the predicted one $\tilde{\mathcal{C}}$ is defined as follows:

$$NMI(\tilde{\mathcal{C}}, \mathcal{C}) = \frac{I(\tilde{\mathcal{C}}, \mathcal{C})}{(H(\tilde{\mathcal{C}}) + H(\mathcal{C}))/2} \tag{7}$$

with

$$I(\tilde{\mathcal{C}}, \mathcal{C}) = \sum_k \sum_j P(\tilde{c}_k \cap c_j) \log \frac{P(\tilde{c}_k \cap c_j)}{P(\tilde{c}_k) P(c_j)}$$

and

$$\begin{aligned} H(\tilde{\mathcal{C}}) &= -\sum_k P(\tilde{c}_k) \log P(\tilde{c}_k) \\ H(\mathcal{C}) &= -\sum_k P(c_k) \log P(c_k) \end{aligned}$$

The Davies-Boulding index DB is a data intrinsic evaluation measure, which is defined as follows

$$DB = \frac{1}{K} \sum_{i=1}^{n} \max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right) \tag{8}$$

where K is the number of clusters, C_k is the centroid of cluster, σ_k is the average distance of all elements in cluster k to centroid c_k, and $d(c_i, c_j)$ is the distance between centroids i and j. The lower is this DB index value, the better is the clustering since this corresponds to low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity).

7. Experiments

On the basis of the categorized comparable corpora collected from Wikipedia, we assess the benefit of mixing native similarity measures with comparability on a 1-NN classification task and on a k-medoid clustering (Kaufman & Rousseeuw, 1987) (Kaufman & Rousseeuw, 1990) task.

7.1 1-NN classification task

We first study the effect of mixing similarity and comparability on the 1-NN classification task while varying the parameter $\alpha \in [0, 1]$.

Figures 3 and 4 show that the similarity/comparability mixing has a significant impact for the two variants C_{VA_1} and C_{VA_2} since it allows reducing by 3% the error rate of the classification for the English language documents and 1.5% for the French language documents. However, comparatively, the C_{LG} measure improves slightly for both languages the classification accuracy, and is less stable when α varies.

7.2 k-medoids clustering task

We study here the effect of mixing comparability and similarity measures on a k-medoids clustering task for all three comparability measures. We used the previously defined AC, NMI and DB measures for the assessment of this clustering task.

Figures 5 and 6 show that both AC and NMI measures can be improved up to 15% in the scope of the clustering of French language documents and up to 3% in the scope of the clustering of English language documents for both C_{VA_1} and C_{VA_2} measures. However, once again, the C_{LG} brings comparatively less improvement for both languages.

Figure 7 depicts the DB measure as a function of parameter α, for all three comparability measures. It is shown that, for C_{VA_1} and C_{VA_2}, this ratio decreases for some *good* α values, especially for the French language, whereas for the measure C_{LG}, this value increases in general. A *good* mixing of the comparability and similarity measures has thus a positive impact when using C_{VA_1} and C_{VA_2} measures and a rather negative impact when using the C_{LG} measure.

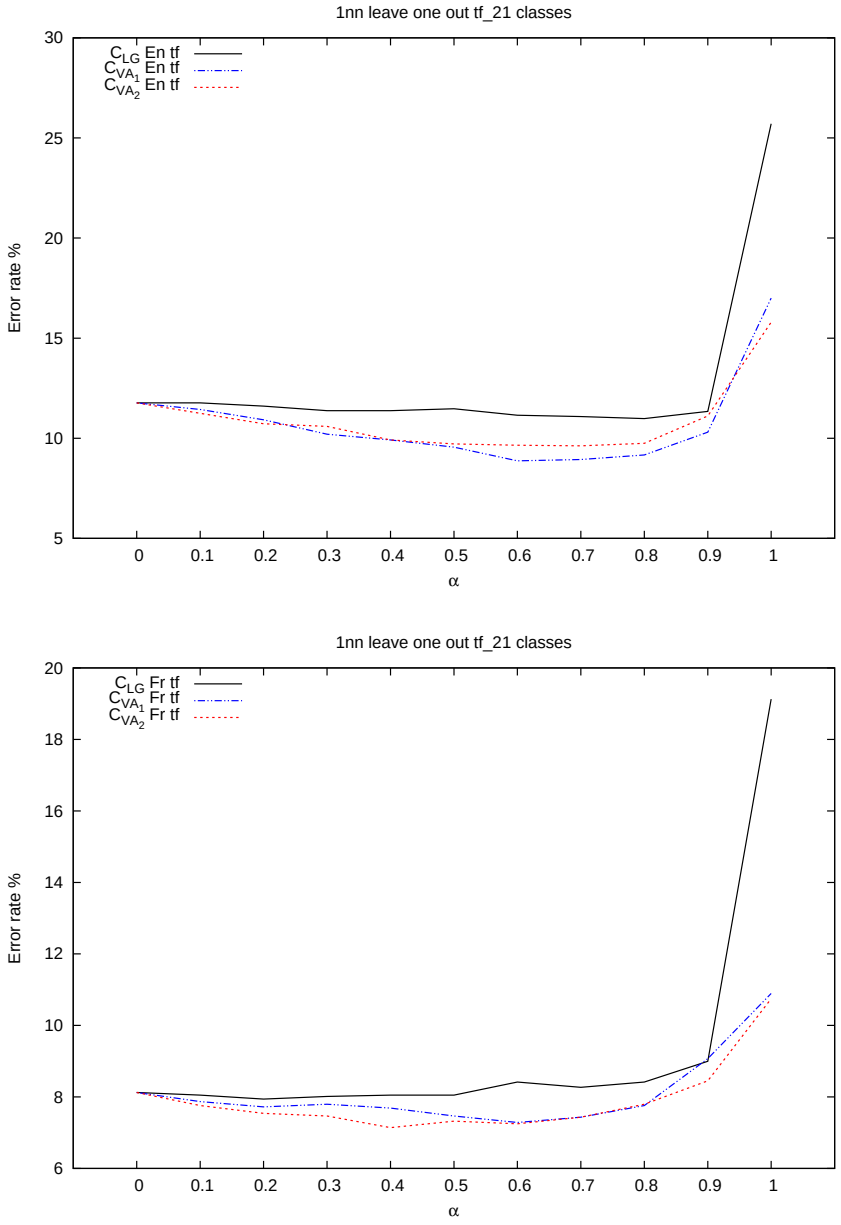

Figure 3: Comparability/similarity mixing effect on the 1-NN classification task, according to the leave one out error rate (top EN documents, bottom FR documents). C_{LG} (black plain curve), C_{VA_1} (blue dashdotted curve), C_{VA_1} (red dotted curve) measures are given as a function of the mixing parameter α.

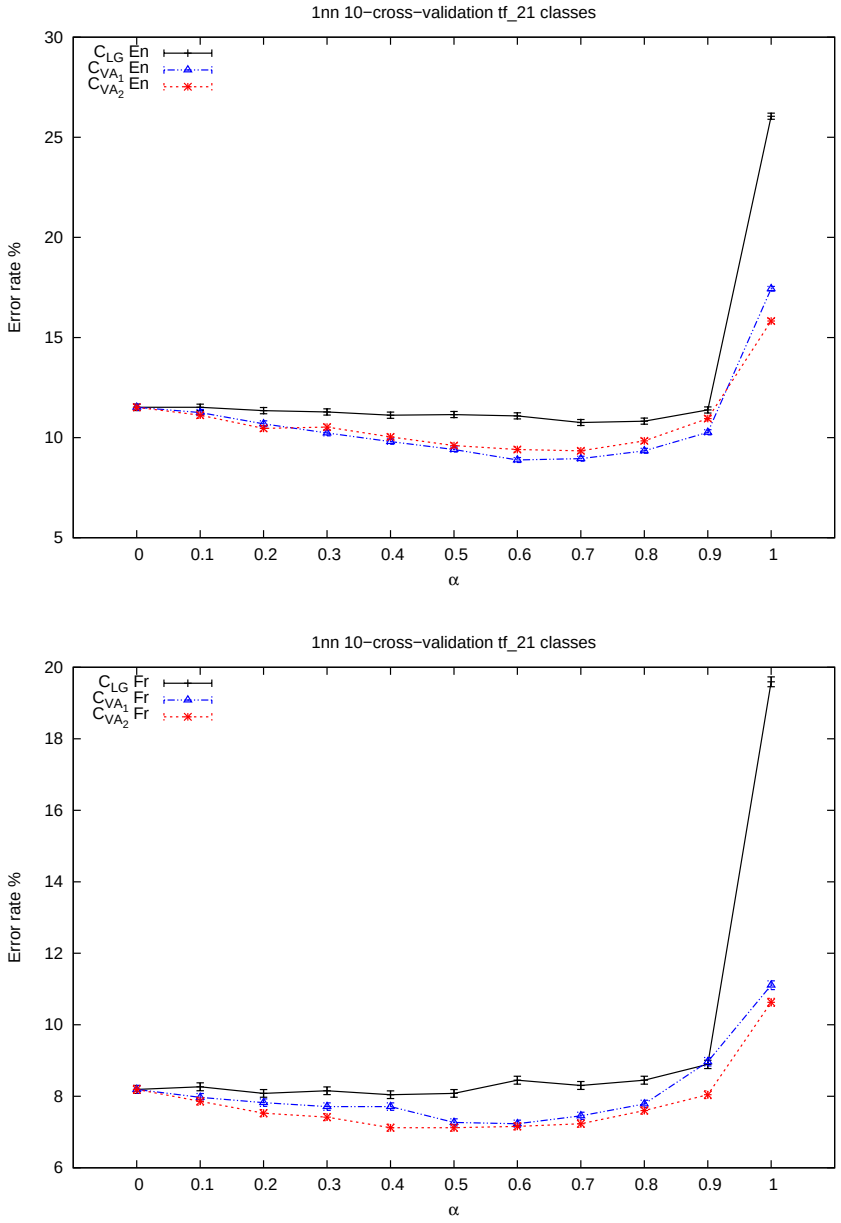

Figure 4: Comparability/similarity mixing effect on the 1-NN classification task, according to 10-fold cross validation error rate (top EN documents, bottom FR documents). C_{LG} (black plain curve), C_{VA_1} (blue triangle dashdotted curve), C_{VA_1} (red star dotted curve) measures are given as a function of the mixing parameter α.

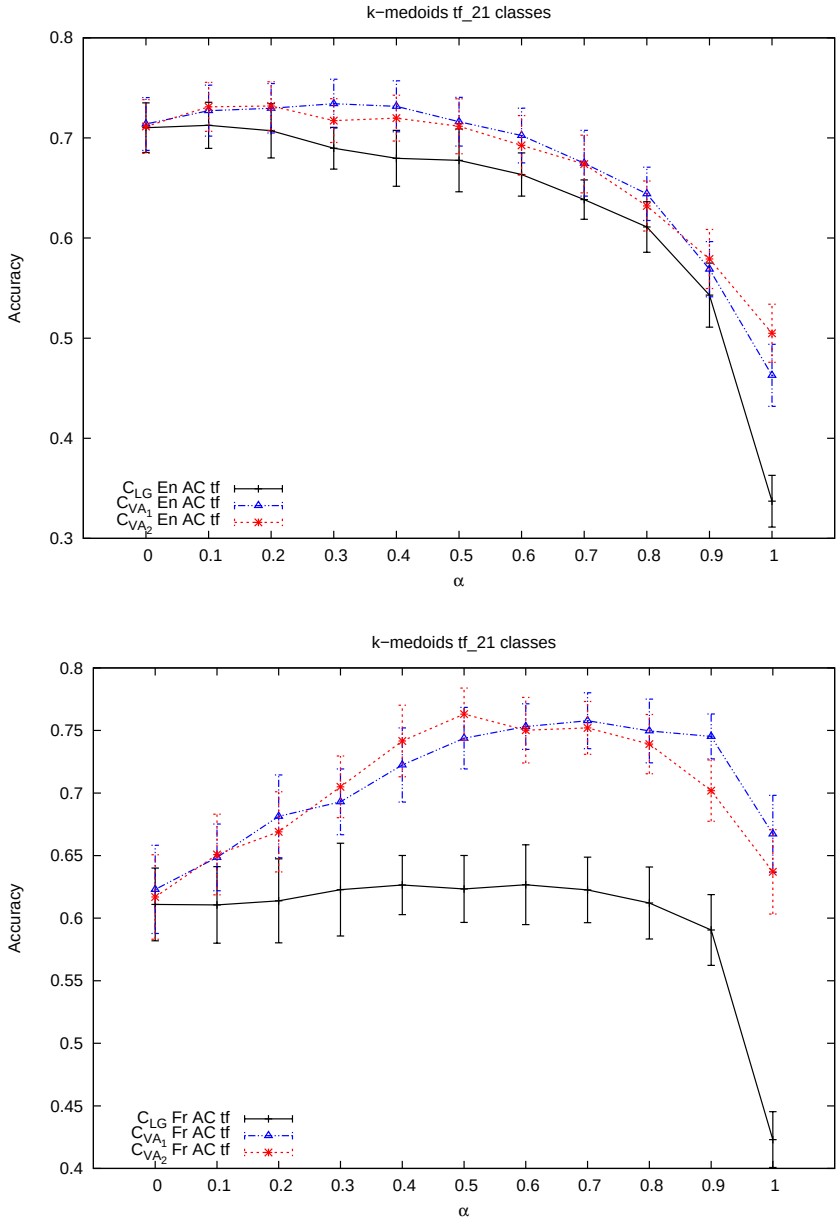

Figure 5: Evaluation of the comparability/similarity mixing on the k-medoids clustering accuracy (AC) (top EN documents, bottom FR documents). C_{LG} (black plain curve), C_{VA_1} (blue triangle dashdotted curve), C_{VA_1} (red star dotted curve) measures are given as a function of the mixing parameter α.

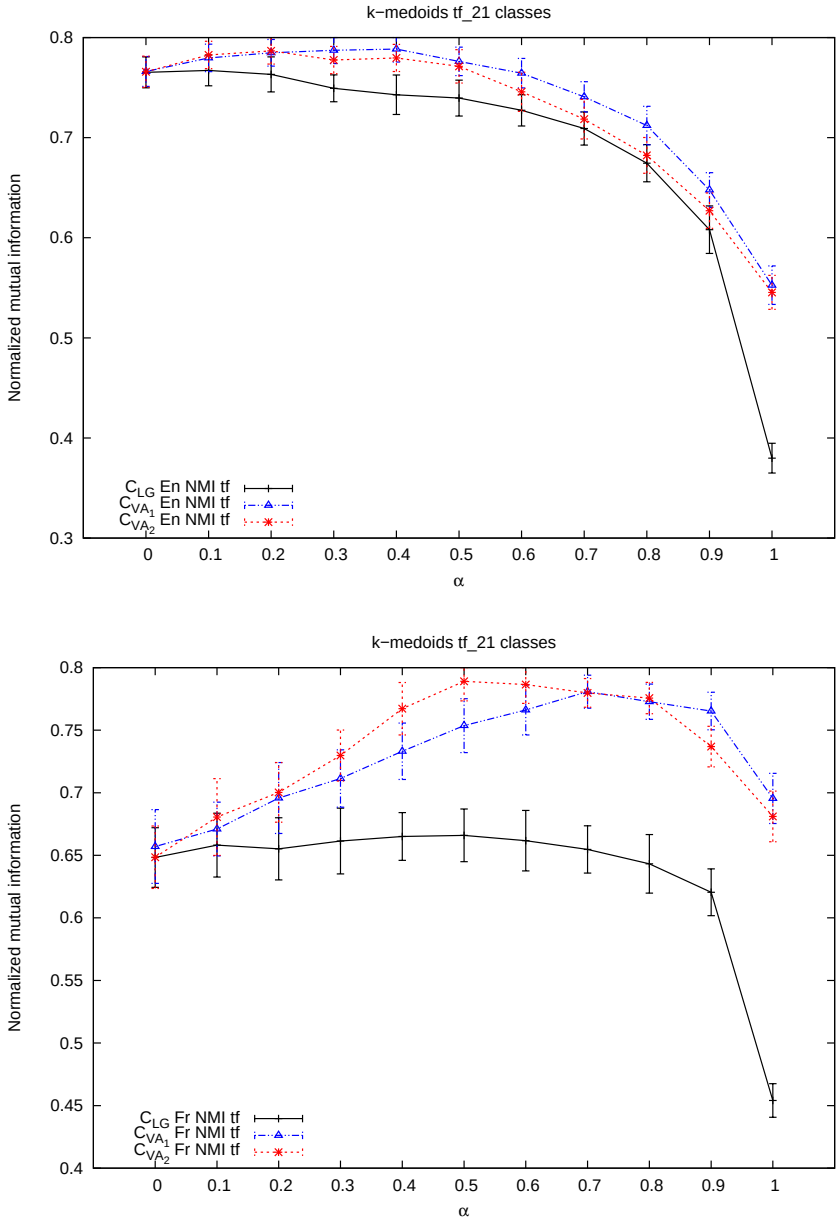

Figure 6: Evaluation of the mixing of comparability and similarity measures on the k-medoids clustering according to the NMI measure (top EN documents, bottom FR documents). C_{LG} (black plain curve), C_{VA_1} (blue triangle dashdotted curve), C_{VA_1} (red star dotted curve) measures are given as a function of the mixing parameter α.

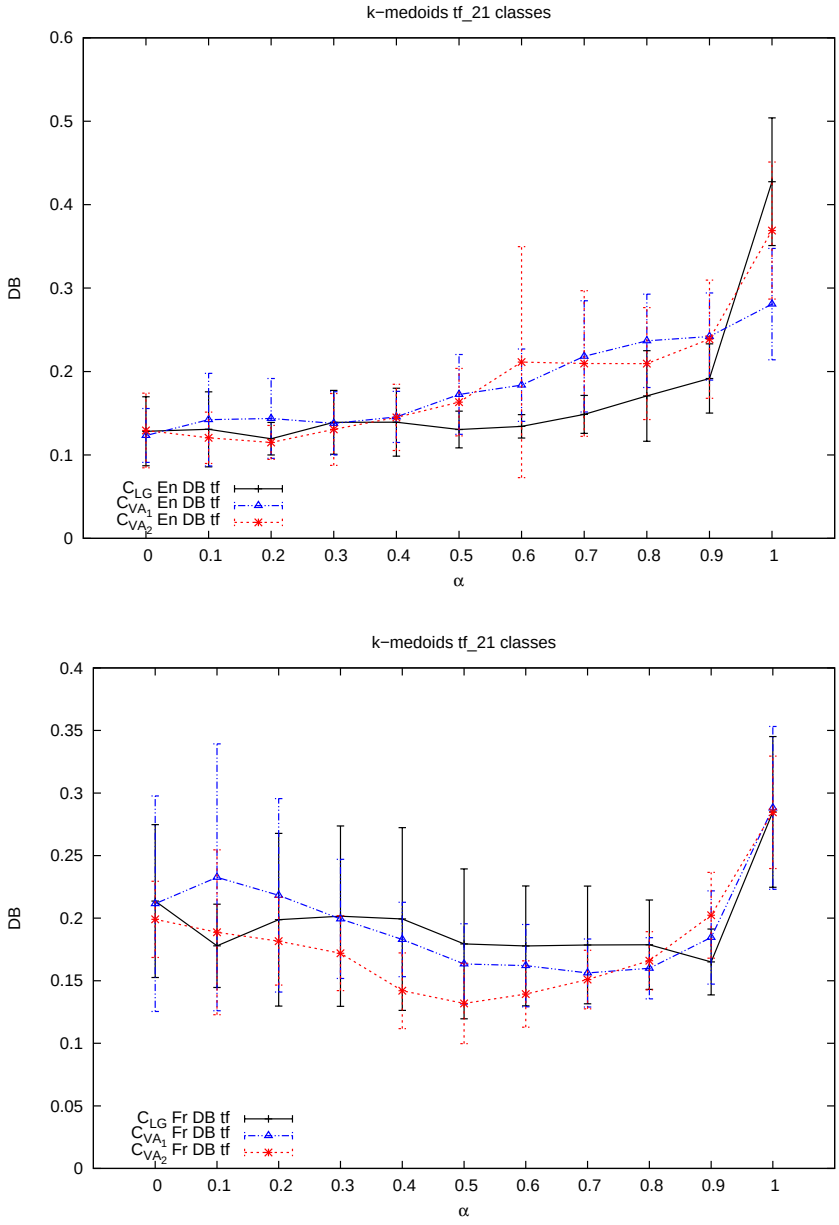

Figure 7: Comparability/similarity mixing effect on a k-medoids clustering according to the DB measure (top EN documents, bottom FR documents). C_{LG} (black plain curve), C_{VA_1} (blue triangle dashdotted curve), C_{VA_1} (red star dotted curve) measures are given as a function of the mixing parameter α.

8. Analysis and conclusions

In this paper, we have proposed a new approach for the co-clustering and co-categorization of bi-lingual data when a comparability mapping exists. This approach, that could be characterized as a kind of three-mode clustering or categorization, is based on the concept of similarity *induced* by a comparability bipartite graph. The three-mode data analysis scheme is implemented as a mixing model used to merge *native* and *induced* similarity measures inside each of the two linguistic space. The assessment of this mixing model on purely synthetic random data is quite informative and demonstrates the noise reduction capability of the method.

On real bilingual textual data, the approach involves a quantitative comparability measure that is based on the exploitation of a bilingual dictionary. To this end, two variants of the comparability measure proposed by (Li & Gaussier, 2010) have been proposed to adapt this measure to clustering and categorization tasks. The implementation of our model on semi-manually constructed comparable corpora collected from the Wikipedia resource shows to be quite effective. Our detailed experimentation shows that the mixing of *native* similarity measures with a quantitative comparability measure has a clear impact on the classification and clustering accuracies. It is noticeable that the improvement is more important in the French linguistic space comparatively to the English linguistic space. Furthermore, our approach works specifically well for the C_{VA_1} and C_{VA_2} comparability variant measures with stable and robust classification or clustering result improvements.

It nevertheless has a small positive impact when the C_{LG} measure is used, leading to conclude that taking into account of the frequency of occurrence of lexical entries and frequencies of their translations into the comparability measure design is of crucial importance for thematic classification or clustering of bilingual English/French documents. One potential explanation is that these frequencies of occurrence pair well with the *tf-idf* heuristic that takes place in native *cosine* similarity. Moreover, according to our results, the choice of the value of the mixing parameter α is quite important. A relatively high α value (between 0.5 and 0.8), that slightly favors the *induced* similarity measures comparatively to the *native* similarity, will be a good choice in general. Finally our experimentation shows that the C_{VA_2}, whose symmetrization is homogeneous to an arithmetic mean, is more robust than C_{VA_1}, a result that needs to be consolidated on other independent experiments.

In terms of perspective, ensuring the scalability and generalizing the approach and experimentation are major prospects to help constructing thematic comparable corpora on demand.

The bilingual dictionary is a particularly important resource in our approach, since the quality of the comparability mapping linking the two linguistic spaces directly relies on it. The impact of the coverage of the dictionary relatively to the corpus has been partly studied in (Li & Gaussier, 2010) and (Ke et al., 2014). In the context of comparable thematic data processing, it is likely that the enrichment of a general bilingual resource by introducing domain specific terminology entries would bring some benefit.

Finally, another perspective is to expand it to various pairing of languages for which bilingual resources are available, in particular bilingual dictionaries.

Acknowledgements

This work has been partially funded by the French National Research Agency (ANR-METRICC).

References

Amini, M.-R., & Goutte, C. (2010). A co-classification approach to learning from multilingual corpora. *Machine Learning*, *79*(1-2), 105–121.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead.. In *In Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*.

Davies, D. L., & Bouldin, D. W. (1979). A cluster separation measure. *Pattern Analysis and Machine Intelligence, IEEE Transactions*, *PAMI-1(2)*, 224–227.

Déjean, H., & Gaussier, E. (2002). Une nouvelle approche a l'extraction de lexiques bilingues à partir de corpus comparables. *Lexicometrica*, *Numéro spécial, corpus alignés*, 1–22.

EAGLES (1996). Expert advisory group on language engineering standards guidelines: http://www.ilc.pi.cnr.it/eagles96/browse.html. Tech. rep., EAGLES.

Fung, P., & Yee, L. Y. (1998). An ir approach for translating new words from nonparallel, comparable texts. In *Proc. of the 36th ACL meeting, Vol. 1*, ACL '98, pp. 414–420, Stroudsburg, PA, USA. ACL.

Jagarlamudi, J., Daumé, III, H., & Udupa, R. (2011a). From bilingual dictionaries to interlingual document representations. In *Proc. ACL-HLT - Vol. 2*, HLT '11, pp. 147–152, Stroudsburg, PA, USA. ACL.

Jagarlamudi, J., Udupa, R., Daumé, III, H., & Bhole, A. (2011b). Improving bilingual projections via sparse covariance matrices. In *Proc.s of the Conf. on EMNLP*, pp. 930–940, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kaufman, L., & Rousseeuw, P. J. (1987). *Clustering by means of Medoids, in Statistical Data Analysis Based on the L_1–Norm and Related Methods*. North-Holland.

Kaufman, L., & Rousseeuw, P. J. (1990). *Finding groups in data: an introduction to cluster analysis*. John Wiley and Sons, New York.

Ke, G., Marteau, P.-F., & Ménier, G. (2014). Variations on quantitative comparability measures and their evaluations on synthetic French-English comparable corpora. In *LREC 2014, the 9th edition of the Language Resources and Evaluation Conference*, p. pp, Reykjavik, Iceland.

Li, B., & Gaussier, E. (2010). Improving corpus comparability for bilingual lexicon extraction from comparable corpora. In *COLING*, pp. 644–652.

Li, B., Gaussier, E., & Aizawa, A. (2011). Clustering comparable corpora for bilingual lexicon extraction. In *Proc. of the 49th ACL-HLT- Vol. 2*, pp. 473–478, Stroudsburg, PA, USA. Association for Computational Linguistics.

Marteau, P.-F., & Ménier, G. (2013). Similarités induites par mesure de comparabilité : signification et utilité pour le clustering et l'alignement de textes comparables.. In *TALN*, p. 515–522.

Mirkin, B. (1996). *Mathematical Classification and Clustering*. Kluwer Academic Publishers.

Munteanu, D. S., Fraser, A., & Marcu, D. (2004). Improved machine translation performance via parallel sentence extraction from comparable corpora. In *HLT-NAACL*, pp. 265–272.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, *18*(11), 613–620.

Schmid, H. (1994). Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of the Int. Conf.e on New Methods in Language Processing*, pp. 44–49.

Schmid, H. (2009). TreeTagger, www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/..

Spärck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*, *28*, 11–21.

Van Mechelen I, Bock HH, D. B. P. (2004). Two-mode clustering methods:a structured overview. *Statistical Methods in Medical Research*, *13(5)*, 363–394.

Vu, T., Aw, A. T., & Zhang, M. (2009). Feature-based method for document alignment in comparable news corpora. In *Proceedings of the 12th EACL Conf.*, pp. 843–851, Stroudsburg, PA, USA. ACL.

Wei Xu, X. L., & Gong, Y. (2003). Document clustering based on non-negative matrix factorization. In *SIGIR'03*, pp. 267–273.